國立臺灣大學電機資訊學院生醫電子與資訊學研究所

碩士論文

Graduate Institute of Biomedical Electronics and Bioinformatics
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

利用腸道微生物菌相早期檢測大腸癌和大腸腺瘤的新型機器學習方法

A novel machine learning pipeline for early detection of colorectal cancer and colorectal adenoma using gut microbiome data

廖乃勳

Nai-Shun Liao

指導教授：莊曜宇 博士
陳佩君 博士

Advisor: Eric Y. Chuang, Sc.D., EMBA
Pei-Chun Chen, Ph.D.

中華民國 112 年 7 月

July 2023

doi:10.6342/NTU202303093

誌謝

在進行碩士研究的這段時間裡，我得到了許多人的指導與幫助，讓我能夠順利完成這份碩士論文。首先，我要特別感謝我的指導教授，莊曜宇教授以及陳佩君教授。感謝莊老師在我開始碩士研究時就依據我大學時期的經歷安排了這份碩士論文的研究題目，並在開會時提供給我許多非常寶貴的意見，使這份研究能夠順利的完成。感謝陳老師在指導時不厭其煩與我討論研究進展，並在遇到問題時給予我許多非常實用的建議，使我能夠修正研究上遇到的問題。此外，我也要感謝賴亮全教授、蔡孟勳教授、盧子彬教授、李建樂教授與陳翔瀚教授，在開會時或是私底下都能教導我許多非常寶貴的建議並在研究上給予我許多幫助，讓我在進行研究的過程中能夠持續的進步，順利的完成這份研究。

除了師長外，我也要感謝實驗室的學長姐。感謝源懋學長在我開始碩士研究時就非常盡心地教導我許多腸道菌相的研究方法與知識，並且在研究生涯上給予我許多實用的建議，讓我在研究上能夠順利地進行。感謝易展學長教導我許多大腸癌相關的知識，並且以臨床上的經驗提供給我許多研究上的建議，讓我在研究上能夠更貼近臨床的需求。感謝 Nam 學長在研究上給予我許多幫助，在模型架構上能夠給我許多建議。感謝佳興學長、采蓉學姊、羽辰學姊、韋霓學姊與兆榮學長，在學業、研究以及實驗室生活中都給予我非常多的建議，在我需要協助時也不吝嗇地給予幫助，讓我能夠更快融入實驗室當中。也要感謝實驗室的同學，家妤在研究上的積極態度在不知不覺中讓我向她看齊；翰儒在研究與學業上真的給我許多幫助；冠緯總是能解答我各式各樣的疑問。同學們的互相鼓勵與幫助讓我在研究上教學相長。還要感謝實驗室的學弟妹：于瑄、育明、庚昀與姜麟，在研究以及實驗室生活上都能一起互相交流。在實驗室的大家庭與學長姐、同學、學弟妹們一起努力進步、一起互相交流鼓勵，這段碩士班研究的相處時光讓我的研究生涯留下難忘的回憶。

最後我要感謝我的家人們。感謝家人們在我就讀碩士班的期間照顧與關心我的生活所需、傾聽我的想法，在學業上也不斷的鼓勵與支持，我才能夠順利的完成碩士班的學習並且完成這份研究。

doi:10.6342/NTU202303093

摘要

大腸直腸癌（簡稱大腸癌）在美國與台灣皆是第三大診斷癌症。通過大腸癌篩檢和診斷可以找出高風險的患者並且大幅降低大腸癌的長期風險。許多研究已經表明大腸癌與腸道微生物菌相之間存在許多關聯。利用機器學習模型來檢測潛在患者的腸道菌相有潛力比傳統的大便篩檢測試更早地檢測到大腸癌。在這篇研究當中，我們構建了一個新的機器學習流程，使用微生物菌相數據來識別大腸癌、大腸腺瘤和健康組別，並評估每個人的大腸癌風險分數。從 SRA 數據庫或其他研究中提供的數據中收集了具有 16S rRNA 定序數據的糞便樣本。根據 ANCOM-BC 演算法和卡方檢定，共識別出 109 個與大腸癌相關的菌屬。使用 10 組交叉驗證對隨機森林分類器進行訓練並且通過外部驗證資料評估模型的分類表現。結果顯示，在區分對照組和大腸癌組方面，隨機森林模型具有優異的分類性能，在 10 組交叉驗證中有 90%的 AUC 並在外部驗證中有 82%的 AUC。在通過分類對照組對比腺瘤加大腸癌組以達到大腸腺瘤早期篩檢的策略中，隨機森林模型在 10 組交叉驗證中表現出 87%的靈敏度，在外部驗證中表現出 97%的靈敏度。最後使用 ANCOM-BC 演算法找出的 7 個生物標記菌屬被用來計算微生物風險得分（MRS），可以被用來作為大腸癌的風險指標。總而言之，我們開發了一種使用 16S rRNA 腸道微生物菌相數據的 CRC 分類新流程，並識別出了特定於大腸癌的腸道微生物菌屬。該流程和生物標記菌屬可以作為早期檢測 CRC 的非侵入性工具使用。

關鍵字：大腸直腸癌、腸道菌相、機器學習、微生物風險得分、糞便早期篩檢

doi:10.6342/NTU202303093

Abstract

Colorectal cancer (CRC) is the third leading diagnosed cancer and cause of cancer death in the United State and Taiwan. The long-term risk of CRC can be managed through the identification of high-risk patients by CRC screening and diagnosis. Many studies have shown the associations between CRC and gut microbiome. The machine learning models have the potential to detect CRC earlier than the conventional stool screening test. We constructed a novel machine learning pipeline to identify CRC, colorectal adenoma, and healthy groups, and evaluated the risk of CRC for each person using microbiome data. Stool samples with16S rRNA sequence data were collected from the NCBI SRA database or supplementary data provided in studies. In total, 109 CRC-associated genera were identified based on ANCOM-BC algorithm and chi-square test. Random forest (RF) classifiers were training with 10-fold cross validation (CV). Model performance was evaluated by the external validation. Our results showed that the RF model illustrated excellent performance with 90% AUC for 10-fold CV and 82% AUC for external validation in classifying control vs CRC groups. RF model performed well with 87% sensitivity for 10-fold CV and 97% sensitivity for external validation in early detection strategy by classifying control vs adenoma plus CRC groups. Finally, 7 biomarkers identified by ANCOM-BC algorithm were utilized to calculate a microbial risk score (MRS), which could be regarded as an index the possibility of CRC. In summary, we developed a new pipeline for CRC classification using 16s rRNA gut microbiome data and identified CRC-specific gut microbiome genera. The pipeline and biomarkers could be used as a non-invasive tool for the early detection of CRC.

Key word: CRC, Gut microbiome, Machine learning, MRS, Stool-based screening

doi:10.6342/NTU202303093

Contents

doi:10.6342/NTU202303093

doi:10.6342/NTU202303093

List of tables

doi:10.6342/NTU202303093

List of figures

doi:10.6342/NTU202303093

doi:10.6342/NTU202303093

Chapter 1. Introduction

1.1 Colorectal cancer

Colorectal cancer (CRC) is a disease characterized by uncontrolled proliferation of abnormal cells specifically affecting the colon or rectum [1]. CRC is the third leading diagnosed cancer and cause of cancer death in the United State [2] and Taiwan [3]. A substantial portion of CRC cases and fatalities can be attributed to modifiable risk factors such as smoking, an unhealthy diet, excessive alcohol consumption, physical inactivity, and obesity [4].

The majority of CRCs originate from noncancerous growths known as polyps, which develop in the inner lining of the colon or rectum [1]. One type of polyps called adenomas are recognized as precursors of CRC [5]. Adenomas may slowly progress to CRC over time by growing through the mucosa and invading blood or lymph vessels. Having higher risk of developing CRC, advanced adenomas is a kind of adenoma defined as an adenoma greater than 1 cm, more than three adenomas of any size, or an adenoma with villous histology [6]. Patients with advanced adenomas are more likely to be diagnosed with CRC.

Because of the slowly progression from adenoma to CRC, a significant proportion of CRC incidence can be prevented through the adoption of regular screening practices and surveillance [6]. Visual examination and stool-based test are two major methods for current CRC screening strategies. Visual examinations, including colonoscopy and flexible sigmoidoscopy, have the best performance for CRC screening. However, visual examinations require bowel cleaning and invasive surgery. Stool-based tests, such as

doi:10.6342/NTU202303093

guaiac-based fecal occult blood test (gFOBT) and Fecal immunochemical test (FIT), are low cost and non-invasive but with low sensitivity for adenoma and advanced adenoma. Both of the screening methods have their advantages and limitation.

1.2 Colorectal cancer and gut microbiome

The Gut microbiome is a large population of microorganisms that lives in the host digest system and involves in the host nutrition, metabolism, and immunity. Dysbiosis of the human gut microbiome plays a significant role in the development of CRC through several different mechanisms. First, microorganisms such as *F. nucleatum* can induce cell inflammation and activate signaling pathway to promote tumor development on intestinal epithelial cell. Second, microbial metabolisms produce metabolites from dietary and host compounds. Secondary bile acids and short chain fatty acids (SCFA) are highly correlated to gut microbiome and diets. These metabolites can influence the risk and formation of tumor. Third, toxins such as Cytolethal distending toxin (CDT) and colibactin produced by gut microbiome can cause DNA damaging effects. These genotoxin can induce double strand DNA degrade and cause genomic instability [7].

Wong and Yu [8] indicated that the two major potential clinical applications in CRC are detecting the screening/prognostic biomarkers and modulation for CRC treatment/prevention. Due to the limitation of current fecal immunochemical test, the stool-based gut microbial genes and metabolites have the potential to be the non-invasive screening or prognostic biomarkers for adenoma, advanced adenoma and CRC. Besides, modulating the gut microbiome can reduce the adverse effects and mediate the anticancer effects of immunotherapy and chemotherapy. Using dietary intervention, gut microbial

doi:10.6342/NTU202303093

probiotics, prebiotics, fecal microbiota transplantation (FMT) and other methods, gut microbiome can be modulated to prevent CRC or improve treatment response of CRC. The findings from several studies have presented an opportunity to apply gut microbiome discoveries into practical applications that can significantly reduce the incidence and mortality rates of CRC.

1.3 Prokaryotic 16S rRNA gene sequencing

The prokaryotic 16S rRNA gene is about 1500 bp long, encoding the small subunit ribosomal RNA fragment of prokaryotic ribosomes. The structure of 16S rRNA gene includes highly conserved gene regions and nine hypervariable regions, called V1 ~ V9 regions. The highly conserved regions allow universal primer binding across different bacteria and archaea, and the hypervariable regions can be used to classify taxon and explore the taxonomic composition of microbiome, as shown in **Figure 1.1** [9]. The most commonly used combination of hypervariable regions for 16S rRNA sequencing are V1 ~ V2/V3, V3 ~ V4 and V4 regions [9].

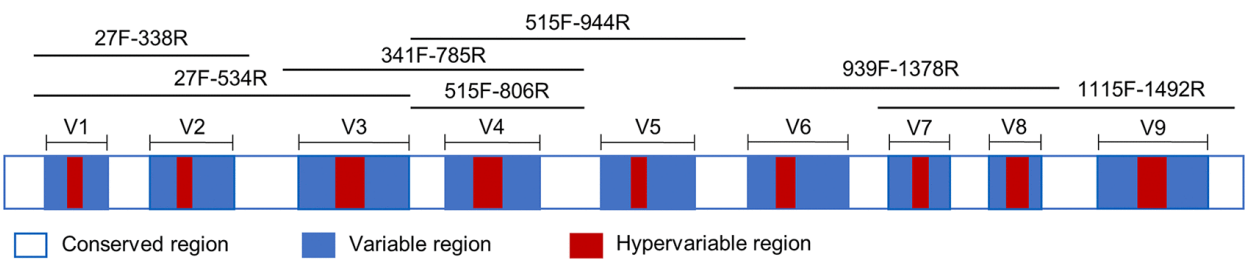


Figure 1.1: The structure of prokaryotic 16S rRNA gene and different primer pairs to sequence hypervariable regions [9].

The 16S rRNA sequencing and analysis pipeline start with sample preparation. Stool, tissue, soil or water samples can be collected and preprocessed to extract the environmental DNA. Polymerase chain reaction (PCR) is used to amplify these DNA

doi:10.6342/NTU202303093

samples. After sequencing these amplicons, the sequencing machine can generate fastq file containing sequencing reads of nucleotides and quality for each base pairs.

Based on the amplicon sequencing data analysis pipeline of Quantitative Insights Into Microbial Ecology 2 (QIIME 2) [10] (**Figure 1.2** [10]), the first step is to demultiplex raw sequencing reads in the fastq file by detecting the barcodes and mapping them back to their samples. Next, the denoising and clustering step uses denoising methods such as DADA2 [11] to remove low quality and chimeric reads based on quality scores, correct amplicon sequencing errors in reads and join denoised paired-end reads. After denoising and clustering, Amplicon sequence variant (ASV) feature table and representative reads for each ASV are generated. The ASV feature table contains sequence counts for each ASV in samples, which is called absolute abundance. Relative abundance is calculated by applying total-sum scaling (TSS) to absolute abundance. The representative reads are selected from each ASV to reduce the number of reads in further analysis steps. To identify the potential organisms each ASV represented, Taxonomy is assigned to each ASV by comparing the representative read of ASV to a 16S rRNA reference database, like SILVA [12] and Greengenes [13].Finally, the ASV table and taxonomy assignment result can be applied to several different analyses, such as diversity analysis and differential abundance analysis. Diversity analysis includes alpha- and beta-diversity analyses. Shannon, Simpson and observed index are common alpha-diversity indices to measure the richness and evenness of ASV in each sample. Beta diversity like Bray-curtis and unifrac compare the similarity or dissimilarity between two samples. Differential abundance analysis, such as Linear Discriminant Analysis Effect Size (LefSe) [14] or Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) [15],

doi:10.6342/NTU202303093

identify significant ASVs which are higher or lower abundant between two group of samples.

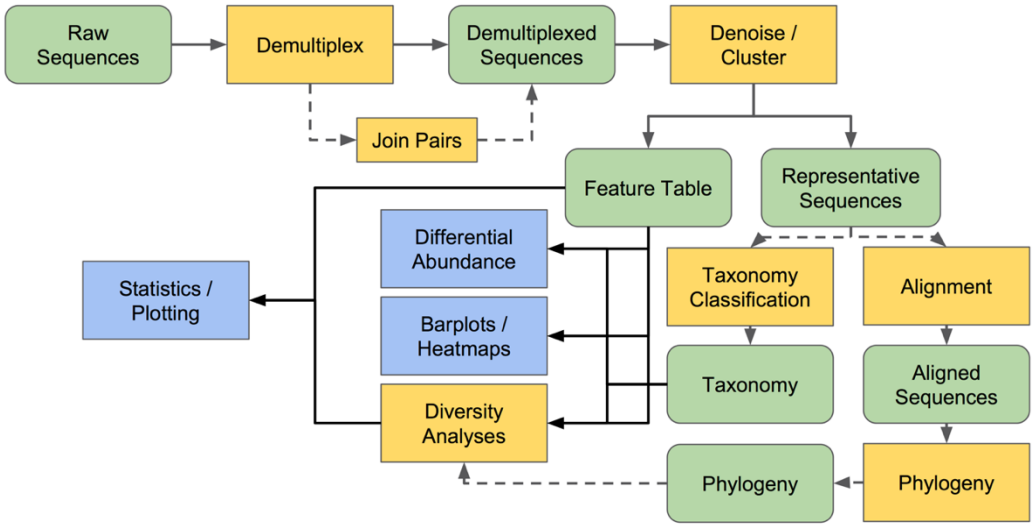
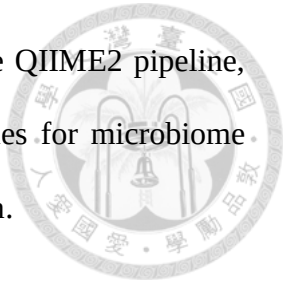


Figure 1.2: QIIME2 amplicon sequencing data analysis pipeline [10].

1.4 Easy Microbiome Analysis Platform (EasyMAP)

Easy Microbiome Analysis Platform (EasyMAP) [16] is an online platform for 16S rRNA gene sequencing data analysis. The analysis pipeline of EasyMAP is based on QIIME2 pipeline described in **Section 1.3**. In brief, the raw reads in the sequence files are demultiplex into the samples based on the sample metadata uploaded by user. The DADA2 algorithm is conducted for quality control and denoising. The taxonomy classifiers are used to do the taxonomy assignment. Pretrained V3-V4 and V4 classifiers on Greengenes and Silva are provided. Classifiers for specific regions can also be trained based on the primer set provided by user. For data visualization and analysis, alpha diversity plots, beta diversity plots, bar plots and heatmaps are provided in EasyMAP. To

doi:10.6342/NTU202303093



further use the ASV table and taxonomy assignment result from the QIIME2 pipeline, EasyMAP also integrates LefSe and PICRUSt [17] analysis modules for microbiome differential abundance analysis and functional composition prediction.

The advantages of the EasyMAP against QIIME2, Mothur [18] or other analysis tools are online platform, easy file management, user-friendly interface, and step-by-step guidance. EasyMAP analysis is conducted on an online web server, eliminating the need for users to prepare their own computing resource. User can manage their file and output results on the EasyMAP web page, intead of a command line interface Linux server. The user-friendly interface and step-by-step guidance provided by EasyMAP enable users with limited bioinformatics knowledge to navigate through the analysis conveniently. Users with limited bioinformatics knowledge can follow the instruction on the EasyMAP web page and the tutorial to complete the whole pipeline. Therefore, EasyMAP is a user-friendly tool specifically designed for conducting comprehensive analysis of 16S rRNA sequencing data.

1.5 Motivation

The motivation of this study is the potential clinical applications of the gut microbiome. The gut microbiome had proven to affect various human disease and cancer by affect disease progression and prognosis, and had the potential to act as the screening biomarker for disease prevention or the probiotics for disease treatment, including colorectal cancer. Colorectal cancer (CRC) ranks as the third most commonly diagnosed cancer and significantly contribute to cancer-related death in both the United States and Taiwan. Despite its impact, the long-term risk of CRC can be effectively managed

doi:10.6342/NTU202303093

through the identification of high-risk patients and underwent CRC screening and diagnosis, followed by regular colonoscopy surveillance [19]. Colonoscopy surveillance is the most precise way to identify both adenoma, advanced adenoma and CRC, but it is expensive, non-convenient and invasive. The stool-based CRC screening tools, such as FIT and gFOBT, are low-cost and non-invasive, but they have low sensitivity to adenoma and advanced adenoma and may produce false-positive test results. Therefore, this study is aimed at develop classification methods based on stool microbiome data for non-invasive CRC screening test in early detection. Additionally, a scoring model is also built to indicate the risk of getting CRC for patients. Patients with higher score may have higher risk of getting CRC, and can get further surveillance by colonoscopy.

doi:10.6342/NTU202303093

Chapter 2. Materials and methods

This study is consisted of two parts. The first part is to construct of classification model, which could be used as a screening model for CRC prevention using gut microbiome data. The second part is to establish a scoring model, which could give a score to potential CRC patients that indicate the risk for getting CRC. The flow chart of this study is in **Figure 2.1**. Stool samples with 16S rRNA gene sequencing data were collected from published studies (**Section 2.1**). 16S rRNA gene sequencing data were downloaded from NCBI SRA database and preprocess by EasyMAP pipeline. The ASV tables generated after preprocessing were merged and convert to abundance and appearance data (**Section 2.2**). Biomarker selection method was based on ANCOM-BC and chi-square test introduced in **Section 2.3** and **Figure 2.2.** Identified biomarker features were served as the models input of random forest (RF) machine learning model and microbial risk score (MRS) model for adenoma and CRC. The training and validation method of the RF model was described in **Section 2.4**. The MRS model for adenoma and CRC was constructed based on MRS framework described in **Section 2.5**.

doi:10.6342/NTU202303093

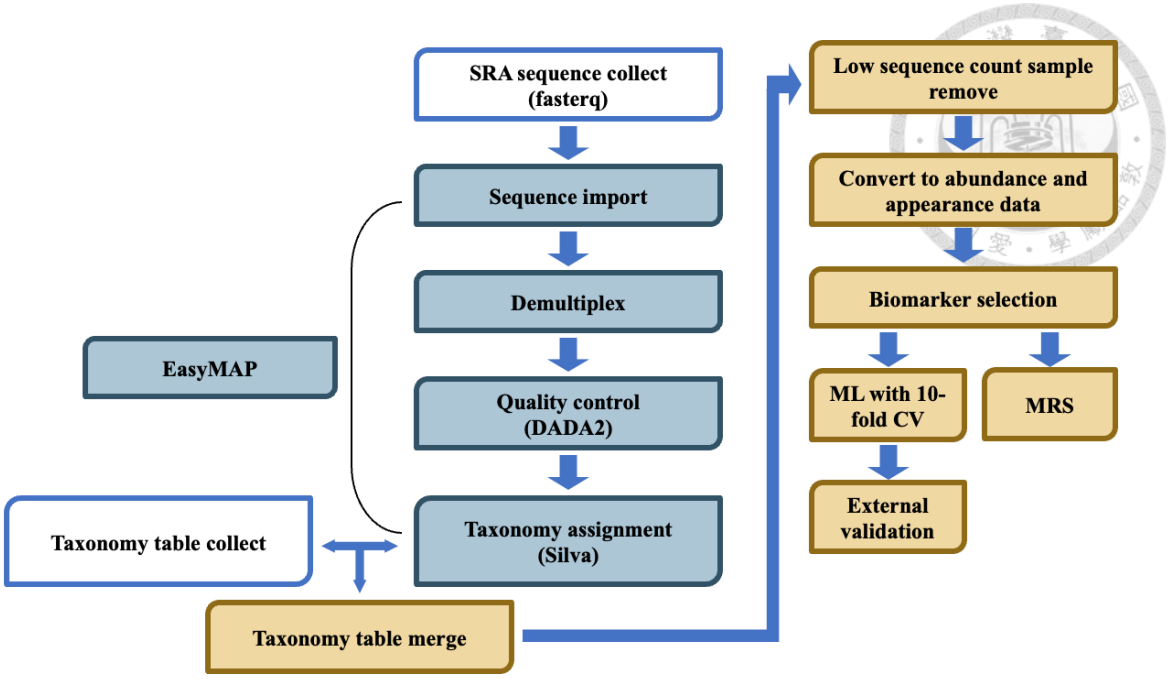


Figure 2.1: The workflow overview of the data preprocessing and model training.

2.1 Published datasets collection

Published studies sequencing stool-based 16S rRNA sequencing data from patients with CRC, advanced adenoma (AA), adenoma and healthy controls prior to treatment or colonoscopy were included in this study. Patients were diagnosed by colonoscopy. Patients without adenoma or CRC were included for healthy controls. Metadata that provide labels and information of the samples were collected from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database or supplementary data provided by datasets. 16S rRNA gene sequencing fastq files of samples in the datasets with data uploaded to NCBI SRA database were accessed and downloaded using SRA access code of each dataset and fasterq-dump tool in SRA toolkit. In the datasets providing 16S rRNA gene sequencing data in supplementary data, fastq files of samples were directly downloaded. Comma-separated values (CSV) ASV tables

doi:10.6342/NTU202303093

of samples in the datasets providing ASV tables were directly downloaded from supplementary data. Samples were excluded if metadata were missing.

There were five datasets collected in this study. Reducing the effect of different country on gut microbiome, the RF and MRS models were constructed by the datasets from USA. In Baxter dataset, patients were inclued in four cities in USA/Canada, able to collect 58 mL of blood and a stool sample. In Dadkhah dataset, patients were included for clinical trial from January 2014 and June 2015. In Zackular dataset, patients were included in four cities in USA/Canada, able to collect 58 ml of blood two times and complete an gFOBT kit. In Yang dataset, patients were included by colonoscopy examination at Tongji University Affiliated Tenth People's Hospital from January 2014 to September 2014. In Cong dataset, patients were included from the affiliated hospital of Qingdao University.

For the RF models, Baxter and Dadkhah datasets were selected to train the models with different classification strategies. For the MRS model, the Baxter dataset was the discovery cohort to classify the control and CRC groups. The RF and MRS models were external validated by the Zackular dataset.

doi:10.6342/NTU202303093

Table 2.1: The study design of the published datasets.

Study	Data storage	Country	Sequencing region
Baxter [20]	SRA	USA/Canada	V4
Dadkhah [21]	SRA	USA	V1-V3
Zackular [22]	Fastq files in supplementary	USA/Canada	V4
Yang [23]	ASV table in supplementary	China	V3-V4
Cong [24]	SRA	China	V3-V4

2.2 Data preprocessing

Fastq files of 16S rRNA sequencing data contained raw reads and their sequencing quality score. Raw reads were mixed with reads from different samples, had uneven quality and lacked of taxonomy assignment. Therefore, data preprocessing for 16S rRNA sequencing data was needed to reveal the gut microbiome information contained in raw reads data.

Raw sequence data were uploaded to EasyMAP website (http://easymap.cgm.ntu.edu.tw/) and processed by EasyMAP pipeline (**Figure 1.2**). Raw sequencing reads were demultiplexed to their belonging samples and trimmed from the left sides to remove the PCR primer. Based on the quality plot, reads were truncated from the right sides to remove the low-quality end. Paired-end reads were kept in enough length to allow paired-end joining. DADA2 method was used to quality control, join paired-end

doi:10.6342/NTU202303093

reads and denoise reads into ASVs. Taxonomy of each ASV was assigned by the classifiers training with the specific sequencing regions based on Silva database. ASV tables from different datasets were merged by assigned taxon of ASVs. The relative abundance data was obtained by the Total-Sum scaling (TSS) applying to the absolute abundance ASV table. The appearance data were transformed from the abundance data by denoted the observed ASVs in *k*th sample as 1 and unobserved ASVs as 0.

After data preprocessing, samples from different datasets were merged into three tables. One was the sample metadata, which contained sample identifiers, sample labels and study indices. Another table was ASV table with abundance data, which contained sample identifiers, ASVs and their assigned taxon and abundance for each sample in each ASV. The other table was ASV table with appearance data, which was the same ASV table structure but with appearance data.

To study and visualize the difference of gut microbiome composition from different groups, The ASV table was analyzed by the phyloseq [25] package (version: 1.42.0) within the Bioconductor R package (version: 3.17). A compositional bar plot was generated to compare the phylum-level composition across different groups. To visualize multidimensional data and maximize information display, principal component analysis (PCA) was conducted, and the first two principal component were plotted.

2.3 Differential abundance analysis and feature selection

After data preprocessing, ASV table contained thousands of ASVs, which most of them were not useful for CRC screening using RF and MRS models. Therefore,

doi:10.6342/NTU202303093

identification of significant ASVs between groups were crucial for improving the performance of RF and MRS models. Differential abundance analysis was performed to identify ASVs that were differently abundant between groups. These significant biomarker features were served as the input of RF and MRS models. As shown in **Figure 2.2,** the ANCOM-BC and chi-square test were used to identify the significant biomarkers from abundance and appearance data. Input features of RF models were determined by combining the biomarkers identified by ANCOM-BC and chi-square test.

2.3.1 Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC)

There are several different differential abundance analysis methods, such as LefSe and Wilcoxon test. These methods only use the relative abundance to identify biomarkers. They do not account for the bias introduced by sampling difference of samples while sequencing.

ANCOM-BC algorithm is a differential abundance analysis that address the bias introduced by unequal sampling fractions [26]. Sampling fraction is defined as the proportion of observed absolute abundance of sample to the unobservable ecosystem. Samples have different sampling fractions due to the different sequencing depth of samples. As shown in **Table 2.2**, the ecosystem A and B is clearly different in absolute abundance, but the sample A and B show the same because of the different sampling fraction between sample A and B. Difference of sampling fractions between samples may cause bias while using observed absolute abundance as the input of differential abundance analysis. Therefore, ANCOM-BC estimate the bias and reduce the false discovery rate.

doi:10.6342/NTU202303093

Based on the MRS framework and the recent review studies [26, 27], ANCOM-BC algorithm, as one of the top-performing differential abundance analysis methods, were recommended to identify the CRC associated ASVs.

Table 2.2: The example of bias introduced by sampling fraction. Samples represent the observed data defined by the library size of sequencing; Ecosystems represent the unobserved data defined by the microbial load in the environments.

	Sample		**Ecosystem**	
	A	**B**	**A**	**B**
ASV1	4	4	12	18
ASV2	2	2	6	9
Sum	6	6	18	27

ANCOM-BC algorithm is a log-ratio linear regression based differential abundance analysis model. First, ANCOM-BC identify ASVs that are systematically absent in a group as structural zeros:

$$\hat{p}_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} I(O_{ijk} \neq 0) \quad (1)$$

i: ASVs (*i* = 1, 2, …, m); *j*: groups (*j* = 1, 2, …, g); *k*: samples (*k* = 1, 2, …, n_j);

p: non-zero proportion of the samples; *O*: observed absolute abundance

If $\hat{p}_{ij} = 0$, then the *i*th ASV is defined as structural zero in the *j*th group. ASVs identified as structural zero among all groups are abandoned in the following steps.

doi:10.6342/NTU202303093

Second, ANCOM-BC estimate the sampling fraction between observed data and unobserved ecosystem. ANCOM-BC assumes the observed absolute abundance data is proportional to the unobserved abundance data in the real microbial ecosystem (**Table 2.2**).

Observed relative abundance:

$$r_{ijk} = \frac{O_{ijk}}{O_{*jk}} \quad (1)$$

Unobserved relative abundance:

$$\gamma_{ijk} = \frac{A_{ijk}}{A_{*jk}} \quad (2)$$

O: observed absolute abundance; *A*: unobserved abundance.

The absolute abundance of a ASV in a random sample is in constant proportion to the absolute abundance in the ecosystem of the sample. This proportional difference between observed and unobserved data is defined as sampling fraction.

$$C_{jk} = \frac{E(O_{ijk}|A_{ijk})}{A_{ijk}} \quad (3)$$

The log transformed absolute abundance data would approximate to normal distribution.

doi:10.6342/NTU202303093

$$y_{ijk} = log(O_{ijk})$$
$$\mu_{ij} = log(\theta_{ij}) \quad (4)$$
$$d_{jk} = log(C_{jk})$$

θ_{ij}: the expected A_{ij} value of *i*th ASV in the *j*th group

In the context where a particular taxon is considered, it is assumed that all subjects, both within and between groups, are independent. In this case, θ_{ij} is regarded as a fixed parameter rather than a random variable. Suppose there are two groups in the samples and for the *i*th ASV, the linear model framework is applied to log-transformed absolute abundance data. The hypothesis can be expressed as follows:

$$y_{ijk} = d_{jk} + \mu_{ij} + \varepsilon_{ijk} \quad (5)$$

$$H_0 : \mu_{i1} = \mu_{i2}$$
$$H_1 : \mu_{i1} \neq \mu_{i2} \quad (6)$$

The difference in the true sample means between the two groups is:

$$\mu_{i1} - \mu_{i2} = E(\bar{y}_{i1} - \bar{y}_{i2}) - (\bar{d}_1 - \bar{d}_2) \quad (7)$$

Under the null hypothesis $\mu_{i1} = \mu_{i2}$, $E(\bar{y}_{i1} - \bar{y}_{i2}) - (\bar{d}_1 - \bar{d}_2) \neq 0$, unless $\bar{d}_1 = \bar{d}_2$. Due to the presence of differential sampling fractions, which are specific to each sample, the numerator of the standard t-test under the null hypothesis for these microbiome data is non-zero. For the equation 7, the first two sections $\mu_{i1} - \mu_{i2}$ and $E(\bar{y}_{i1} - \bar{y}_{i2})$ should be normal distribution. Therefore, the sampling fraction bias $(\bar{d}_1 - \bar{d}_2)$ can be estimated by the method proposed in the ANCOM-BC methodology. The

doi:10.6342/NTU202303093

ANCOM-BC algorithm control the false discovery rate and perform great testing power compared to other differential abundance analysis method [26].

The ASV tables containing absolute abundance data were subjected to identify biomarkers between Control, Adenoma and CRC groups using the ANCOM-BC package (version: 2.0.3) implemented in Bioconductor R package (version: 3.17). The abundance data were analyzed by the global test of ANCOM-BC across the control, adenoma and CRC groups. The log fold change of the identified significant ASVs were compared between control and CRC groups.

2.3.2 Feature selection

In many studies, features for model input were commonly selected by differential abundance analysis methods, or even by RF model. In this study, combining both ANCOM-BC and chi-square test, two types of significant biomarkers were selected. These two methods were selected to identify the biomarkers from abundance and appearance data.

Two types of biomarkers were selected from abundance or appearance data. Biomarker selected by ANCOM-BC using abundance data shows significant difference between groups. Most of the ASVs were low abundant in samples and couldn't distinguish the difference between groups while using abundance data. The appearance data take account for the difference of observed or unobserved rather than high or low abundance. The appearance data representing the ASVs that observed in samples were transformed from the abundance data by denoted the observed ASVs in *k*th sample as 1 and unobserved ASVs as 0. The appearance ratio of the ASVs between Control, Adenoma

doi:10.6342/NTU202303093

and CRC groups were analyzed by chi square test to identify significant biomarkers that highly observed in the specific groups of samples.

Finally, these two types of significant biomarkers (p-value < 0.01 in ANCOM-BC and/or chi-square test) were selected and combined together. The abundance data with feature selected based on the combined biomarkers were utilized as the input of machine learning model. The combined biomarkers from two test can address the significance from two different type of data. The abundance data selected based on biomarkers identified by ANCOM-BC were utilized as the input of the microbial risk score calculation (**Figure 2.2**).

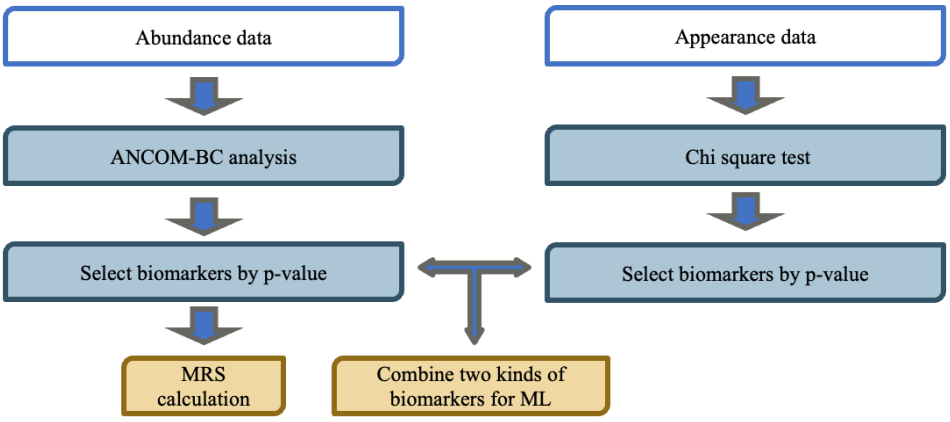


Figure 2.2: Feature selection method for abundance and appearance data.

2.4 Machine learning model training and external validation

Machine learning and deep learning are often built for classification using high-dimensional data. Due to the limitation of sample amount, random forest machine learning method was built to reduce overfitting and handle the non-linear features.

doi:10.6342/NTU202303093

Random forest (RF) is a kind of machine learning algorithm. RF ensemble multiple decision trees that construct by nodes and leaves. Each node represents a test that split the data into branches, and branches end in each leaf represents a class label. By assembling the output of all the trees, RF reduces overfitting problem, handles both categorical and continuous variables and doesn't affect by non-linear features. Three main hyperparameters of RF are number of trees, minimum number of samples to split a node, and the number of features to consider for a split. These hyperparameter may affect model performance. Therefore, they can be tuned by grid search and cross validation.

The evaluation metrics of the RF models included in this study were accuracy, AUC, sensitivity and specificity. Accuracy is the ratio of the number of correct predictions with the number of total samples. AUC is the area under Receiver operating characteristic (ROC) curve that presents the correlation between true positive rate and false positive rate of the model predictions. Sensitivity, also called true positive rate or recall, represents the percentage of positive samples that are correctly predicted as positive. Specificity, also called true negative rate, is the percentage of negative samples that are correctly predicted as negative. These four metrics can evaluate the model performance for classification.

To classify the control, adenoma and CRC groups by the stool microbiome data, the random forest classification models were construct using the scikit-learn machine learning package (version: 1.2.2). The input features were selected by ANCOM-BC and chi-square test (**Figure 2.2**). For the classification models that had external validation dataset, the random forest classifiers were training and grid searching for best hyperparameters with stratified 10-fold cross validation. Due to the unbalanced datasets

doi:10.6342/NTU202303093

across three groups, each class of the RF models were weighted inversely proportional to class frequencies in the datasets. The weighted classes could reduce the effect of the unbalanced datasets. The accuracy, AUC, sensitivity and specificity metrics, which were widely used for classification models, were used to evaluate the model performance and select the best classifier. After 10-fold cross validation, the optimized model was validated by an external validation dataset to avoid overfitting and generally evaluate the ability of the stool microbiome random forest classifiers across different datasets, such as technical differences in microbial data generation. For the classification models without external validation dataset, models were validated by stratified 10-fold cross validation.

2.5 Microbial risk score (MRS)

Microbial risk score (MRS) [28] is aimed at summarizing the disease-specific microbial profiles into a continuous risk score, which can be employed to assess and predict diseases susceptibility. MRS framework is inspired by the polygenic risk score (PRS). In recent years, PRS has gained increasing utility in current genomic researches. By integrating the cumulative effect from the risk alleles identified by genome-wide association study (GWAS) into a continuous score, PRS offers a comprehensive and quantitative measure of genetic risk on a disease. However, one primary distinction between MRS and PRS arises from the complex ecosystem of the microbiota, which is driven by interactions among the sub-community of microorganisms or between microorganisms and human host. Therefore, rather than the weighted sum of the relative abundance from the microbial sub-community, the community-based MRS applies

doi:10.6342/NTU202303093

various alpha diversity indexes, which measure the richness and evenness within a single sample, to the sub-community with microbial biomarkers associated with the disease.

The MRS algorithm includes two major steps: (1) sub-community determination; (2) risk score calculation and validation. The first step involves the application of differential abundance analysis, such as ANCOM-BC, ALDEx2 [29] and Maaslin2 [30], on the discovery cohort to identify the ASVs that associate with disease. ANCOM-BC algorithm gives each ASVs a significant p value. Next, the inclusion of ASVs in the sub-community is determined using pruning and thresholding method (P+T method). P+T method prune the ASVs by the p value threshold. ASVs with p value lower than the threshold were included. To determine the p value threshold, P+T method systematically evaluates all possible p value thresholds, ranging from low to high. The lowest p value threshold can include a least three ASVs and the highest threshold can include all ASVs. For each possible p value, one MRS model is constructed on the samples in discovery cohort, which calculated the alpha diversity index (such as Shannon, Simpson and Observed) using ASVs under the threshold. In the next step of the MRS framework, ROC curves were built based on all of the MRS models. The optimal MRS model with the highest Area under ROC curve (AUC) on the discovery cohort is selected. In the second step, the optimal MRS model is calculated on the discovery cohort and validated by the validation cohort.

Based on the MRS framework, the global test of ANCOM-BC was used to identify the significant ASVs across the control, adenoma and CRC groups. The sub-community of biomarkers that maximized the mean MRS value difference between control and CRC groups was determined by pruning and thresholding method. The Shannon alpha diversity

doi:10.6342/NTU202303093

index were used to calculate the MRS value on both discovery and validation cohort. The statistical significance of average MRS between groups were tested by unpaired two-tailed Student's t test.

doi:10.6342/NTU202303093

Chapter 3. Results

In this study, three datasets from USA containing stool samples and corresponding 16S rRNA gene sequencing data were utilized to investigate the potential of gut microbiome biomarkers as non-invasive screening tool for CRC and adenoma. The overall characteristics of all three datasets were listed in **Section 3.1**. The biomarkers that significant difference across control, adenoma and CRC groups were identified by ANCOM-BC and chi-square test. RF models with different classification strategies were trained and external validation with the identified biomarkers. Finally, the MRS model for CRC screening was constructed by the sub-community included 7 bacterial genera, which were selected by ANCOM-BC and P+T method.

3.1 Alterations of gut microbial composition between control, adenoma and CRC groups

In order to identify potential biomarkers in the gut microbiome for the development of a stool-based test for colorectal cancer (CRC), three datasets containing stool samples and corresponding 16S rRNA gene sequencing data were utilized. These datasets consisted of individuals from the United States of America (USA) and/or Canada, including patients diagnosed with adenoma, advanced adenoma, CRC, as well as healthy controls. A comprehensive description of the included datasets can be found in **Table 3.1**. To ensure data quality and consistency, all raw sequencing data underwent preprocessing using the EasyMAP platform. BMI and age were higher in CRC and adenoma groups compared to control groups

doi:10.6342/NTU202303093

Table 3.1: Information of the datasets with stool samples included in the study.

Study	Control (No.)	Adenoma (No.)	Advanced adenoma (No.)	CRC (No.)	Age	BMI
Baxter [20]	187	101	124	127	60.32 (SD=12.18)	27.18 (SD=5.63)
Dadkhah [21]	237	242	73		62.12 (SD=8.74)	27.22 (SD=5.05)
Zackular [22]	30	30		30	58.66 (SD=10.67)	28.18 (SD=5.93)
Total	454	373	197	157		

To visualize the gut microbiome data within the three datasets, a comparison across the control, adenoma, advanced adenoma (AA), and CRC groups was conducted using principal component analysis (PCA) and the relative abundance composition at the phylum level.

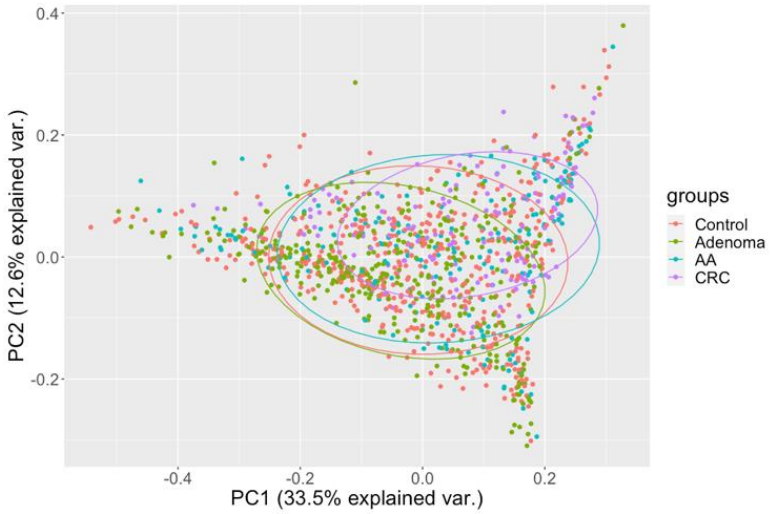


Figure 3.1: Principal component analysis of the gut microbiome samples from all three studies in control, adenoma, advanced adenoma (AA) and CRC groups. The first two PCs each explained 33.5% and 12.6% of variance.

doi:10.6342/NTU202303093

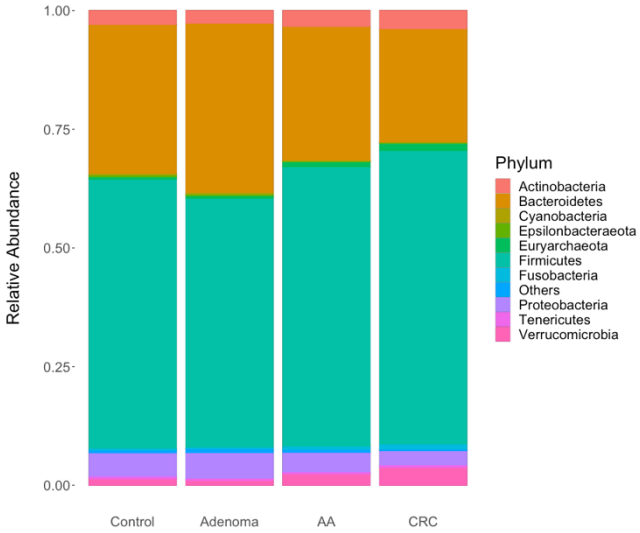


Figure 3.2: Stacked phylum level relative abundance bar plot of the gut microbiome samples from all three studies in control, adenoma, advanced adenoma (AA) and CRC groups. Y axis: the stacked relative abundance of the four groups.

In the principal component analysis plot, CRC groups showed difference compared to control and adenoma groups, where CRC groups was slightly separated with control and adenoma groups in PCA (**Figure 3.1**). The gut microbiome composition of the four examined groups in phylum-level was predominantly characterized by the presence of *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, and *Verrucomicrobia* (**Figure 3.2**). Of particular interest, the *Firmicutes* / *Bacteroidetes* ratio (referred to as the F/B ratio), determined by the abundance of the first two dominant phyla, was relatively higher in CRC group in comparison to the other three groups. With these two results, the gut microbiome data of CRC groups were different with other groups of samples. Besides, due to the lack of validation dataset, the advanced adenoma group was combined to adenoma group in the following feature selection, classification models and MRS model.

doi:10.6342/NTU202303093

3.2 Features selection across control, adenoma and CRC groups

In order to conduct a more comprehensive analysis of the gut microbiome variations between the control, adenoma, and CRC groups, the log fold changes of bacterial genera between the three groups were calculated using the abundance data. As shown in **Figure 3.3,** *Porphyromonas*, *Collinesela*, *Fusobacterium*, *Peptostreptococcus* and *Parvimonas* exhibited significantly higher abundance in the CRC group compared to both the control and adenoma groups. Conversely, the abundance of *Anaerostipes* and *Haemophlius* was found to be comparatively lower in the CRC group compared to the other two groups.

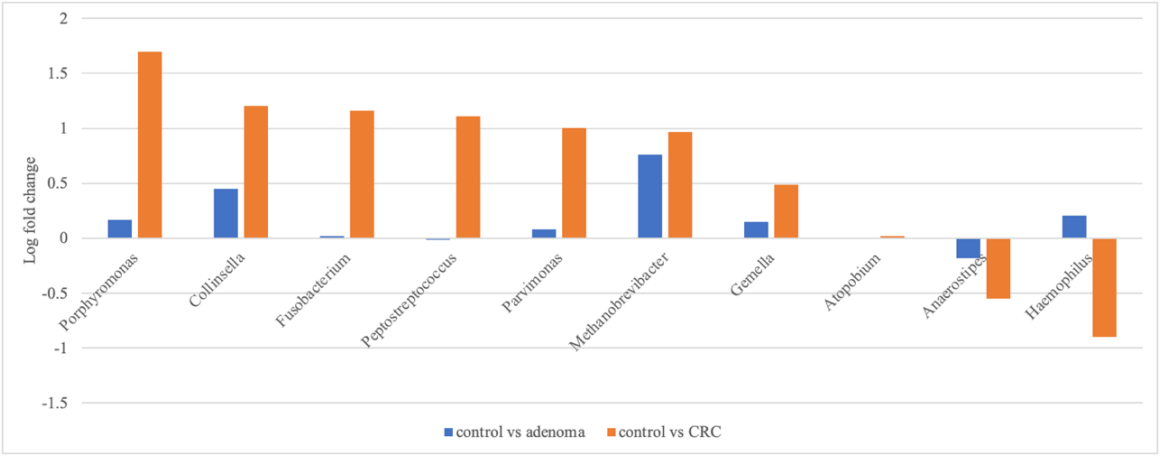


Figure 3.3: Log fold change of the significant bacterial genera among control, adenoma and CRC groups. Y axis: the log fold change in different comparison.

doi:10.6342/NTU202303093

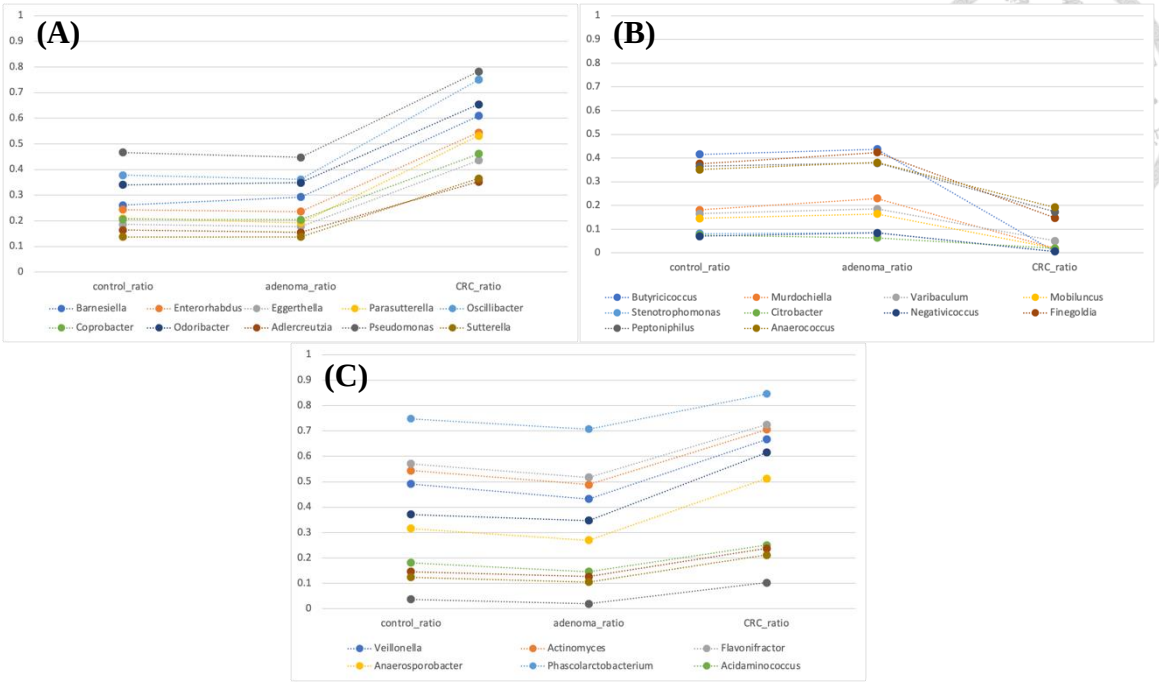


Figure 3.4: Proportions of bacterial genera change in three groups: (A) Lower in CRC group; (B) Higher in CRC group; (C) Lower in adenoma group. Y axis: the percentage of the samples within the group that observed the biomarkers.

Additionally, addressing the challenges posed by the relatively low abundance and sparsity of the ASV table, the appearance ratios of bacterial genera within the three groups were employed. The ASV table with appearance data showed the difference of observed or unobserved in samples rather than high or low abundance. The appearance ratio of a ASV within a specific group was defined as the percentage of the samples within the group that observed the ASV. Notably, a cluster of bacterial genera exhibited a higher occurrence in the CRC group, indicating that these bacterial genera were highly observed in samples from CRC group (**Figure 3.4 (A)**). Samples with these genera observed were more likely to be in CRC group. On the contrary, another cluster of bacterial genera displayed a lower occurrence in the CRC group compared to the other two groups, suggesting they were highly observed in the normal and adenoma groups. Samples without these genera observed may indicated to be in CRC group (**Figure 3.4 (B)**). The

doi:10.6342/NTU202303093

third groups of genera were lower appeared in adenoma group (**Figure 3.4 (C)**), showing that there were still differences between adenoma and other two groups.

The high dimensional and sparse ASV table impeded its direct usage for machine learning model prediction. Therefore, the method described in **Figure 2.2** was employed to select gut microbiome biomarkers. Specifically, the genus level ASV table underwent analysis using the global test of ANCOM-BC algorithm and chi-square analysis among control, adenoma and CRC groups. A significance threshold of p-value < 0.01 was applied on both of the testing method, resulting in the selection of 109 biomarkers. 10 biomarkers listed in **Figure 3.3** were selected by ANCOM-BC and 99 biomarkers were selected by chi-square test. These biomarkers were subsequently utilized as input features for the machine learning models.

3.3 Microbial classification models for control, adenoma and CRC groups

Next, to test the ability for using gut microbiome biomarkers selected in **Section 3.2** as CRC stool-based screening tool, random forest classification models with stratified 10-fold cross-validation (CV) were constructed by pooling the Baxter and Dadkhah datasets as training data. Zackular dataset was selected to be the hold-out external validation dataset. The advanced adenoma group was merged with the adenoma group if advanced adenoma group wasn't specified in the classification strategies. The 109 differential biomarkers identified by ANCOM-BC and chi-square test were using as input features.

doi:10.6342/NTU202303093

Table 3.2: Model performance metrics of random forest classifiers in pairwise binary classification strategies. Strategies: negative vs positive class; validation method: 10-fold CV and external validation for each strategy; AUC, accuracy, sensitivity, specificity: the evaluation metrics for models.

Strategy	Validation method	AUC	Accuracy	Sensitivity	Specificity
Control vs Adenoma	**10-fold cross validation**	0.63 (SD=0.05)	0.59 (SD=0.04)	0.79 (SD=0.08)	0.40 (SD=0.09)
	External validation	0.62	0.55	0.73	0.37
Adenoma vs CRC	**10-fold cross validation**	0.90 (SD=0.02)	0.86 (SD=0.05)	0.38 (SD=0.12)	0.97 (SD=0.01)
	External validation	0.84	0.66	0.33	0.98
Control vs CRC	**10-fold cross validation**	0.90 (SD=0.05)	0.85 (SD=0.04)	0.43 (SD=0.13)	0.97 (SD=0.02)
	External validation	0.82	0.67	0.38	0.96

First, the pairwise binary classifications were performed to discover the ability for distinguishing one group over another. Pairwise classifications helped us to find out the pair of groups that were distinctively difference across these three groups. When using the control vs CRC strategy to train and validate the model, the 10-fold CV and external validation AUC were 0.90 and 0.82 separately. The specificity was relatively high in both 10-fold CV and external validation, which were 0.97 and 0.96. However, the accuracy performance drops when using external validation. (**Table 3.2**)

The model performance of adenoma vs CRC strategy showed similar result compare with control vs CRC strategy, with AUC 0.90 and 0.84 in 10-fold CV and external

doi:10.6342/NTU202303093

validation. Nevertheless, the model performance of control vs adenoma strategy showed poor prediction power compare to the other two strategies, with AUC only 0.63 and 0.62 in 10-fold CV and external validation (**Table 3.2**). This result may cause by the lack of distinction between control and adenoma groups in gut microbiome composition **(Figure 3.2)** and appearance proportions of the existed bacterial genera **(Figure 3.4)**.

Adenoma might slowly progress to CRC. Therefore, CRC incidence can be significantly prevented by early detection of adenoma. Early detection of adenoma offers an opportunity for further treatments to prevent the adenoma progression to CRC. For an early detection tool, it is crucial to discover the patients with disease or with higher risk to get disease. The early detection tool with higher sensitivity can identify most of the disease group. The Control vs Adenoma + CRC strategy was trying to detect both adenoma and CRC for early detection of the disease group. The samples in CRC group represented patients that already had cancer, and the samples in adenoma group represented patients that had higher risk to gradually get cancer. Though AUC and accuracy were not as good as the other strategies, this strategy showed higher sensitivity, 0.87 and 0.97, in 10-fold cross validation and external validation (**Table 3.3**). Higher sensitivity represents more patient can be correctly identified, which is important for early detection screening methods. Patients that are identified as adenoma + CRC group can perform further investigation, such as colonoscopy, to verify whether the patients are adenoma or CRC.

doi:10.6342/NTU202303093

Table 3.3: Model performance metrics of random forest classifier in Control vs Adenoma + CRC binary classification strategies.

Strategy	Validation method	AUC	Accuracy	Sensitivity	Specificity
Control vs Adenoma + CRC	10-fold cross validation	0.71 (SD=0.05)	0.66 (SD=0.03)	0.87 (SD=0.03)	0.36 (SD=0.06)
	External validation	0.80	0.72	0.97	0.21

3.4 Microbial risk score for CRC

MRS is a continuous risk score that summarizing the disease-specific microbial profiles. The MRS workflow was conducted in this study to provide a more directly way to assess the risk of patient to get CRC.

To compute the MRS score among control, adenoma and CRC group, the Baxter and Zackular datasets containing both three groups of the samples were included as discovery and validation cohort. Following the MRS workflow, the genus level ASV table with absolute abundance was analyzed by ANCOM-BC algorithm to identify the significant bacterial genera. Then, the sub-community for MRS calculation was determined by P+T method. Specifically, the p-value threshold was determined by the sub-community that maximize the mean difference of the MRS value between control and CRC groups. With p-value threshold < 0.01, 7 bacterial genera were included to calculate the MRS score. The MRS values were computed based on Shannon index. The abundance of the 7 included genera, *Porphyromonas*, *Peptostreptococcus*, *Parvimonas*, *Fusobacterium*,

doi:10.6342/NTU202303093

Haemophilus, *Atopobium* and *Collinsella*, showed significantly different between groups (**Figure 3.3**).

The mean and standard deviation of the MRS score among control, adenoma and CRC groups in Baxter and Zackular datasets were listed in **Table 3.4. Figure 3.5** showed the means and the 95% confidence inetrvals (CI) of MRS among three groups. The differences of means between groups were tested by Student's t test. In Baxter dataset, the average MRS score of the CRC group was significantly higher than the control group ($p = 2.1 \times 10^{-12}$) and the adenoma ($p = 6.9 \times 10^{-7}$) group, and the average MRS score of the adenoma group was significantly higher than the control group ($p = 0.0046$). In Zackular dataset, the average MRS score of CRC group was also significantly higher than the adenoma group ($p = 0.021$). The tests results showed that the means of MRS were different between groups. The means of MRS were increased from control, adenoma to CRC groups, indicated that samples with higher MRS score were related to higher risk of being in the CRC group in these two USA/Canada datasets.

Table 3.4: Mean and standard deviation of the MRS score across control, adenoma and CRC groups in discovery and validation cohort. SE: standard error

Study	Control	Adenoma	CRC
Baxter (discovery)	0.22 (SE=0.025)	0.34 (SE=0.029)	0.62 (SE=0.053)
Zackular (validation)	0.40 (SE=0.10)	0.27 (SE=0.069)	0.56 (SE=0.10)

doi:10.6342/NTU202303093

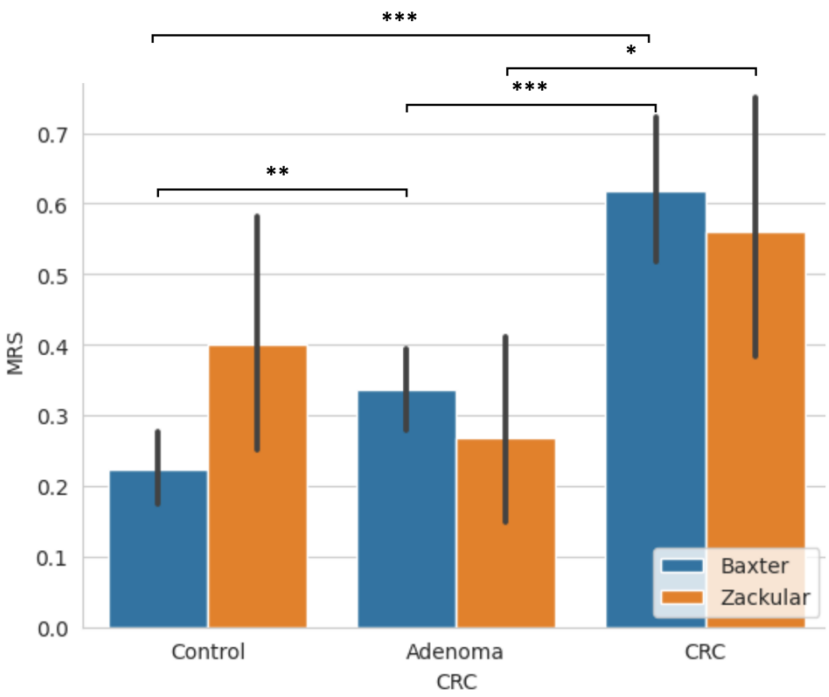


Figure 3.5: Average and 95% CI of the MRS score across control, adenoma and CRC groups in discovery and validation datasets. Y axis: MRS score; X axis: three groups of samples in two datasets; error bar: 95% CI; *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$ (Student's t test).

doi:10.6342/NTU202303093

Chapter 4. Discussion and conclusions

Two kinds of models were built in this study. The results of the random forest model for pairwise and early detection classification showed the potential to classify CRC and adenoma groups using stool based gut microbiome data. In addition, MRS framework was applied in this study. MRS score based on significant biomarkers can give each patients a score that indicate the risk of getting CRC. In our MRS model, only 7 biomarkers were needed in the sub-community to calculate the MRS score, which could be an efficient and low-cost tool for CRC risk evaluation.

In this study, stool-based gut microbiome data from published datasets were processed and compared the difference between control, adenoma and CRC groups. Then, the genus-level biomarkers were identified from abundance and appearance data by combining ANCOM-BC and chi-square test result. These biomarkers were served as the input of RF classification models. Models had great AUC for control vs CRC and adenoma vs CRC strategies with pairwise classification. The early detection strategy, control vs adenoma + CRC, showed great sensitivity that recall most of the patients. Another screening strategy that took account for the advanced adenoma, showed comparable result for sensitivity and specificity with other stool-based screening test.

Most of the included datasets sequenced the V4 or V3-V4 regions of the 16S rRNA gene. Only the Dadkhah dataset sequenced the V1-V3 regions. The difference of the sequencing region might affect the results of the ASV taxonomy assignment. Different data preprocessing method is needed to deal with data in different sequencing regions. Therefore, when performing the taxonomy assignment, the specific taxonomy classifiers

doi:10.6342/NTU202303093

were trained based on the targeted V1-V3, V3-V4 or V4 region by Silva database to precisely assign the taxonomy in the datasets with different sequencing regions. Then, different datasets were merged by the assigned taxon to proceed the further analysis.

The two dominant bacterial phyla in human gut are *Firmicutes* and *Bacteroidetes*. Therefore, the difference of *Firmicutes* / *Bacteroidetes* ratio infer to a huge alteration of bacterial community in gut microbiome. Lots of studies have reported that F/B ratio is related to Obesity [31], Type I diabetes [32] and other diseases. The result of previous studies show that the F/B ratio of the CRC stool samples are higher compared to normal samples [33, 34]. Consistent with previous studies investigating the relationship between CRC and the gut microbiome, our study also indicate that the *Firmicutes* / *Bacteroidetes* (F/B) ratio is relatively higher in the CRC group compared to the healthy control group. This observation aligns with the growing evidence suggesting a potential association between an altered F/B ratio and the development or progression of CRC.

In this study, differential abundance analysis was conducted using the ANCOM-BC algorithm. ANCOM-BC correct the bias introduced by sampling fractions. Different sampling fractions for each sample may cause the observed abundance not representing the real abundance of the unobserved ecosystem. By correcting the sampling fractions, ANCOM-BC control the false discovery rate and perform great testing power. Other differential abundance analysis method, such as ALDEx2 and LefSe are also widely used in microbiome studies. 109 biomarkers on the genus-level, including *Porphyromonas*, *Fusobacterium*, *Peptostreptococcus* and *Parvimonas*, were identified among control, adenoma and CRC groups by the biomarkers selection method (**Figure 2.2**) using the global test of ANCOM-BC and chi-square testing. A group of bacterial

doi:10.6342/NTU202303093

species, such as *Parvimonas micra* [35], *Fusobacterium nucleatum* [36], *Porphyromonas gingivalis* [37] and *Peptostreptococcus stomatis* [37], are widely reported to be associated with development and prognosis of CRC. Despite the limitation and resolution of 16S rRNA gene sequencing, *Porphyromonas*, *Fusobacterium*, *Peptostreptococcus* and *Parvimonas* were also identified in genus-level, which showed similar result with other studies.

There are many types of classification model structure, such as multilayer perceptron (MLP), convolution neuron network (CNN), random forest and support vector machine. We had attempted the deep learning structure, including MLP and CNN, but the AUC performance for control vs CRC strategy were around 0.83, which were lower than the performance of random forest. The reason may be the number of samples were not enough for training a deep learning model.

Among the model performance of the RF in pairwise binary classification, control vs CRC and adenoma vs CRC strategies had the best performance, both had 0.90 in AUC. These results were also higher than other related studies [38, 39], which perform AUC 0.80 and 0.89 in control vs CRC and adenoma vs CRC. Though loss in other evaluation metrics, control vs adenoma strategy still had slightly better sensitivity compared to the other two strategies. Due to similar between control and adenoma groups, while CRC screening, the control vs CRC strategy can be used specific to classify the CRC groups against the other two groups. The sensitivity and the specificity showed different pattern in different model classification strategies. These result might cause by the unbalanced datasets. Though the class weight were balanced by the inversely proportion of class frequencies while model training, the unbalanced dataset

doi:10.6342/NTU202303093

still affected the model performance. Down sampling or other methods is needed to reduce the effect of unbalanced dataset and further improve the model performance.

About the stool-based CRC screening tests, guaiac-based fecal occult blood test (gFOBT) had largely replaced by Fecal immunochemical test (FIT) due to the convenience and effectiveness of FIT [40]. However, FIT still can't perform well on detecting advanced adenoma (AA). Pooled analysis showed that the sensitivity for detecting advanced adenoma is about 23% and the specificity is 94% [41]. Our classification sensitivity for AA + CRC against Control +adenoma was 43% and the specificity was 94%, which was similar with the pooled analysis result of the FIT screening test. This result showed the potential of using gut microbiome for advanced adenoma screening. In fact, PCA analysis (**Figure 3.1**) showed that the advanced adenoma and adenoma groups were different, and the advanced adenoma group was similar with CRC groups. More stool-based gut microbiome datasets with advanced adenoma are needed to validate the result of this study. The FIT screening test is the most widely used method for CRC and advanced adenoma screening. Therefore, our model has the potential to compete with FIT. Furthermore, our model can even combine with FIT result to get a better performance for CRC and advanced adenoma screening.

The MRS framework applied on the Baxter dataset as discovery cohort. The ASVs were pruned using the p value calculated by ANCOM-BC. The sub-community that maximized the average difference of the MRS score between control and CRC groups in discovery cohort. Based on this method, 7 biomarkers were included into the sub-community. The MRS model applied on the Baxter dataset showed significantly increasing MRS score across control, adenoma and CRC group, indicated the ability of

doi:10.6342/NTU202303093

CRC screening using MRS score. Validation on the Zackular dataset showed similar result. The average MRS score of the CRC group was significantly higher than adenoma group in the Zackular datset. Nevertheless, the average MRS score of the adenoma group in Zackular dataset was slightly lower than the adenoma group in Baxter dataset. It might due to the slightly difference in gut microbiome for the adenoma groups between discovery dataset, with both adenoma and advanced adenoma, and validation dataset, without advanced adenoma. Using different alpha diversity index had similar result compared with Shannon index.

To find the potential MRS thresholds for CRC screening, the MRS score of control and CRC groups were considered. As shown in **Figure 4.1,** the MRS score in 70% of samples in control groups were highly accumulate between 0 to 0.25, while only about 40% of the samples in CRC groups were lower than 0.25. Based on the percentiles of two groups (**Table 4.1** and **Figure 4.1**), the potential MRS thrseholds could be set to 0.12 and 0.21, which was the 60 percentile of control group and 40 percentile of CRC group. Samples with MRS score under 0.12, which was lower than the 35 percentile of CRC group, were considered as safe group. Samples with MRS score above 0.21, which neared the 70 percentile of control group, were considered as danger group. Samples with MRS score between 0.12 and 0.21 were considered as warning group.

doi:10.6342/NTU202303093

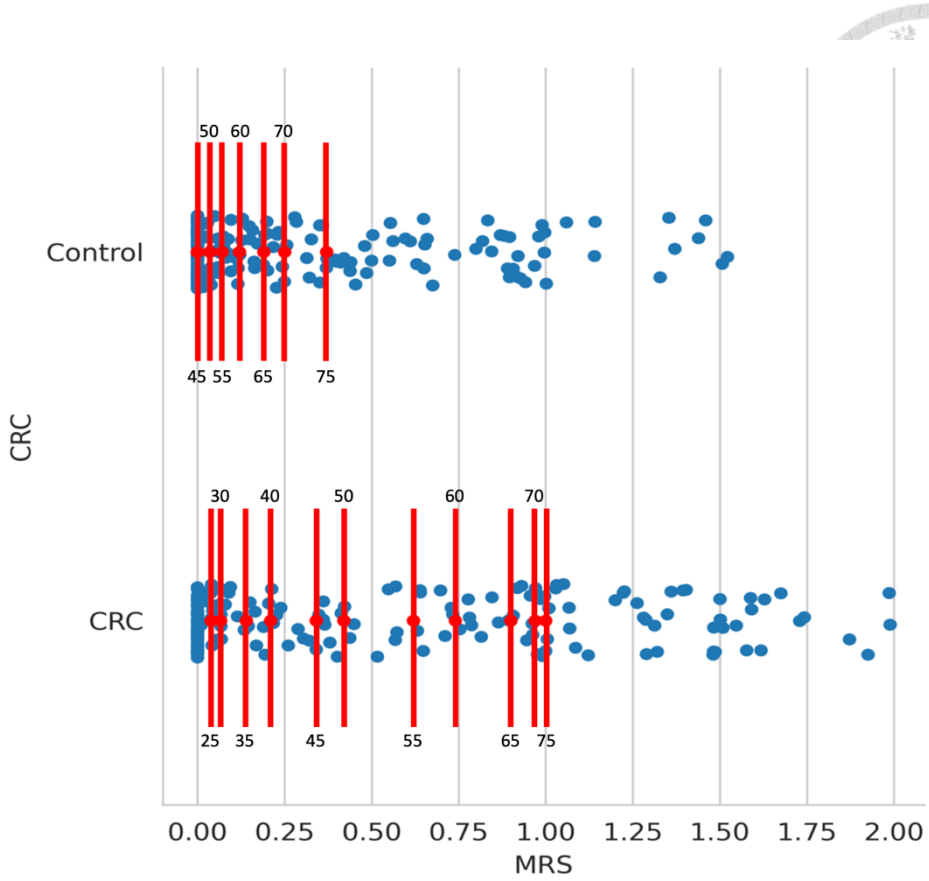


Figure 4.1: Percentiles of the pooled MRS score in control and CRC groups. X axis: MRS score; red bar: the percentiles of MRS score.

Table 4.1: Mean and percentiles of the pooled MRS score in control and CRC groups. SE: standard error

Groups	Mean	Percentiles										
		25	30	35	40	45	50	55	60	65	70	75
Control	0.25 (SE=0.025)	0	0	0	0	0	0.035	0.071	0.12	0.19	0.25	0.37
CRC	0.61 (SE=0.046)	0.039	0.066	0.14	0.21	0.34	0.42	0.62	0.74	0.90	0.97	1.0

Other than the dataset from USA or Canada, studies from different regions may affect the application of the MRS model for CRC screening. Therefore, Yang and Cong

doi:10.6342/NTU202303093

datasets from China (detail information listed in **Table 2.1**) were processed and calculated the MRS score to evaluate the MRS model. Using the same sub-community of 7 biomarkers identified in Baxter discovery cohort, the average MRS score of the two China datasets were higher compared to Baxter and Zackular datasets in both control and CRC groups. Nevertheless, compared between control and CRC groups from the two China datasets, the average MRS score also significantly increased from control to CRC groups (**Figure 4.2** and **Table 4.3**), with $p = 3.4 \times 10^{-6}$ in Yang dataset and $p = 0.0017$ in Cong dataset separately. This result showed the similar pattern with two USA/Canada datasets, which indicated the potential cross-regional application of the proposed MRS model.

Table 4.2: Information of the external validation datasets in China with stool samples

Study	Control (No.)	CRC (No.)	Published year
Yang [23]	50	50	2019
Cong [24]	11	10	2018

Table 4.3: Mean and standard deviation of the MRS score in China datasets. SE: standard error

Study	Control	CRC
Yang	0.70 (SE=0.074)	1.23 (SE=0.075)
Cong	0.51 (SE=0.12)	1.45 (SE=0.23)

doi:10.6342/NTU202303093

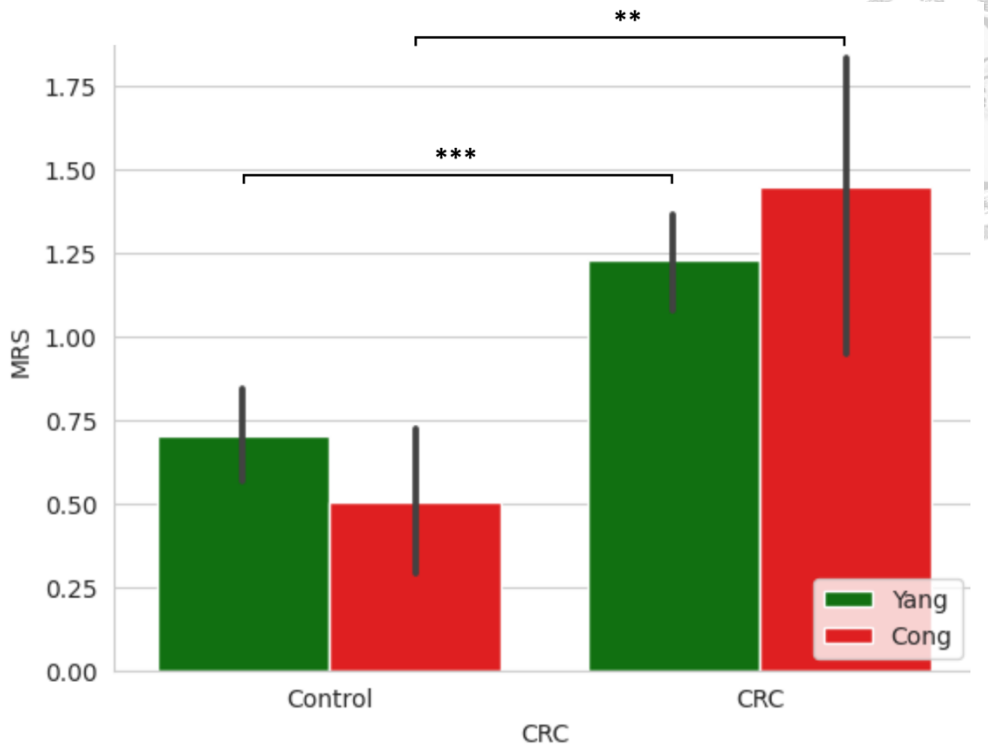


Figure 4.2: Average and 95% CI of the MRS score in China datasets. Y axis: MRS score; X axis: control and CRC groups in two china datasets; error bar: 95% CI. **: $p \leq 0.01$, ***: $p \leq 0.001$ (Student's t test).

To further enhance future research in this study, several aspects could be improved. First, considering the limitations of 16S rRNA gene sequencing data, future research could explore the utilization of more advanced techniques such as shotgun whole genome sequencing (WGS) or 16S rRNA full-length sequencing. These methods offer higher resolution in taxonomy assignment. Besides, WGS also provide additional functional predictions [42]. By incorporating these sequencing approaches, the increased resolution and additional information can potentially improve the accuracy of classification models used in the study and gain more understanding of the gut microbiome composition and its potential function in relation to CRC. Another aspect that can contribute to the improvement of future research is the inclusion of more gut microbiome datasets, specifically those including samples from patients with advanced adenoma. By expanding

doi:10.6342/NTU202303093

the available datasets, the performance the random forest classifier can be enhanced by training and external validation for advanced adenoma classification. The MRS model can also improve to classify different groups of samples, like adenoma or advanced adenoma. The gut microbiome between the adenoma and advanced adenoma (AA) groups showed difference in PCA and phylum-level gut microbiome composition. The advanced adenoma group was even more similar with CRC group in PCA analysis. More datasets containing AA samples is needed to improve the performance of RF models classifying AA groups and construct the MRS models for AA screening. As a continuous score, MRS models can be integrated with other types of data, such as multi-omics data or ages, which can hopefully improve the performance of MRS score.

doi:10.6342/NTU202303093

References

1. *Colorectal cancer facts & figures 2023-2025*. Atlanta: American Cancer Society 2023; Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2023.pdf.

2. Siegel, R.L., et al., *Colorectal cancer statistics, 2023.* CA: A Cancer Journal for Clinicians, 2023. **73**(3): p. 233-254.

3. 衛生福利部. *110 年國人死因統計結果*. 2022 2022/06/30; Available from: https://www.mohw.gov.tw/cp-16-70314-1.html.

4. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA: A Cancer Journal for Clinicians, 2018. **68**(1): p. 31-54.

5. Strum, W.B., *Colorectal adenomas.* New England Journal of Medicine, 2016. **374**(11): p. 1065-1075.

6. Patel, S.G., et al., *Advanced adenomas may be a red flag for hereditary cancer syndromes.* Hereditary Cancer in Clinical Practice, 2021. **19**(1): p. 8.

7. Chattopadhyay, I., et al., *Exploring the role of gut microbiome in colon cancer.* Applied Biochemistry and Biotechnology, 2021. **193**(6): p. 1780-1799.

8. Wong, S.H. and J. Yu, *Gut microbiota in colorectal cancer: mechanisms of action and clinical applications.* Nature Reviews Gastroenterology & Hepatology, 2019. **16**(11): p. 690-704.

9. Abellan-Schneyder, I., et al., *Primer, pipelines, parameters: Issues in 16S rRNA gene sequencing.* mSphere, 2021. **6**(1): p. e01202-20.

10. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.* Nature Biotechnology, 2019. **37**(8): p. 852-857.

doi:10.6342/NTU202303093

11. Callahan, B.J., et al., *DADA2: High-resolution sample inference from Illumina amplicon data.* Nature Methods, 2016. **13**(7): p. 581-3.

12. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Research, 2012. **41**(D1): p. D590-D596.

13. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Applied and Environmental Microbiology, 2006. **72**(7): p. 5069-5072.

14. Segata, N., et al., *Metagenomic biomarker discovery and explanation.* Genome Biology, 2011. **12**(6): p. R60.

15. Lin, H. and S.D. Peddada, *Analysis of compositions of microbiomes with bias correction.* Nature Communications, 2020. **11**(1): p. 3514.

16. Hung, Y.-M., et al., *EasyMAP: A user-friendly online platform for analyzing 16S ribosomal DNA sequencing data.* New Biotechnology, 2021. **63**: p. 37-44.

17. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nature Biotechnology, 2013. **31**(9): p. 814-821.

18. Schloss, P.D., et al., *Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and Environmental Microbiology, 2009. **75**(23): p. 7537-7541.

19. Bjerrum, A., et al., *Long-term risk of colorectal cancer after screen-detected adenoma: Experiences from a Danish gFOBT-positive screening cohort.* International Journal of Cancer, 2020. **147**(4): p. 940-947.

20. Baxter, N.T., et al., *Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions.* Genome Medicine, 2016. **8**(1): p. 37.

21. Dadkhah, E., et al., *Gut microbiome identifies risk for colorectal polyps.* BMJ Open Gastroenterology, 2019. **6**(1): p. e000297.

doi:10.6342/NTU202303093

22. Zackular, J.P., et al., *The human gut microbiome as a screening tool for colorectal cancer.* Cancer Prevention Research, 2014. **7**(11): p. 1112-1121.

23. Yang, Y., et al., *Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer.* Theranostics, 2019. **9**(14): p. 4101-4114.

24. Cong, J., et al., *A pilot study: Changes of gut microbiota in post-surgery colorectal cancer patients.* Frontiers in Microbiology, 2018. **9**: p. 2777.

25. McMurdie, P.J. and S. Holmes, *phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.* PLOS ONE, 2013. **8**(4): p. e61217.

26. Lin, H. and S.D. Peddada, *Analysis of microbial compositions: a review of normalization and differential abundance analysis.* npj Biofilms and Microbiomes, 2020. **6**(1): p. 60.

27. Nearing, J.T., et al., *Microbiome differential abundance methods produce different results across 38 datasets.* Nature Communications, 2022. **13**(1): p. 342.

28. Wang, C., et al., *Microbial risk score for capturing microbial characteristics, integrating multi-omics data, and predicting disease risk.* Microbiome, 2022. **10**(1): p. 121.

29. Gloor, G., *ALDEx2: ANOVA-like differential expression tool for compositional data.* ALDEX Manual Modular, 2015. **20**: p. 1-11.

30. Mallick, H., et al., *Multivariable association discovery in population-scale meta-omics studies.* PLOS Computational Biology, 2021. **17**(11): p. e1009442.

31. Magne, F., et al., *The Firmicutes/Bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients?* Nutrients, 2020. **12**(5).

32. Murri, M., et al., *Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study.* BMC Medicine, 2013. **11**: p. 46.

doi:10.6342/NTU202303093

33. Fang, C.Y., et al., *Colorectal cancer stage-specific fecal bacterial community fingerprinting of the taiwanese population and underpinning of potential taxonomic biomarkers.* Microorganisms, 2021. **9**(8).

34. Wang, Y., et al., *Alterations in the oral and gut microbiome of colorectal cancer patients and association with host clinical factors.* International Journal of Cancer, 2021. **149**(4): p. 925-935.

35. Zhao, L., et al., *Parvimonas micra promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients.* Oncogene, 2022. **41**(36): p. 4200-4210.

36. Osman, M.A., et al., *Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer.* Scientific Reports, 2021. **11**(1): p. 2925.

37. Guven, D.C., et al., *Analysis of Fusobacterium nucleatum, Streptococcus gallolyticus and Porphyromonas gingivalis in saliva in colorectal cancer patients and healthy controls.* Journal of Clinical Oncology, 2018. **36**(15_suppl): p. e15617-e15617.

38. Mo, Z., et al., *Meta-analysis of 16S rRNA microbial data identified distinctive and predictive microbiota dysbiosis in colorectal carcinoma adjacent tissue.* mSystems, 2020. **5**(2): p. 10.1128/msystems.00138-20.

39. Wu, Y., et al., *Identification of microbial markers across populations in early detection of colorectal cancer.* Nature Communications, 2021. **12**(1): p. 3063.

40. Ladabaum, U., et al., *Strategies for colorectal cancer screening.* Gastroenterology, 2020. **158**(2): p. 418-432.

41. Selby, K., et al., *Effect of sex, age, and positivity threshold on fecal immunochemical test accuracy: A systematic review and meta-analysis.* Gastroenterology, 2019. **157**(6): p. 1494-1505.

42. Wirbel, J., et al., *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer.* Nature

doi:10.6342/NTU202303093

Medicine, 2019. **25**(4): p. 679-689.

doi:10.6342/NTU202303093